

(HEART DISEASE PREDICTOR)

Project report submitted in partial fulfillment of the requirement for
the degree of Bachelor of Technology

in

Computer Science and Engineering

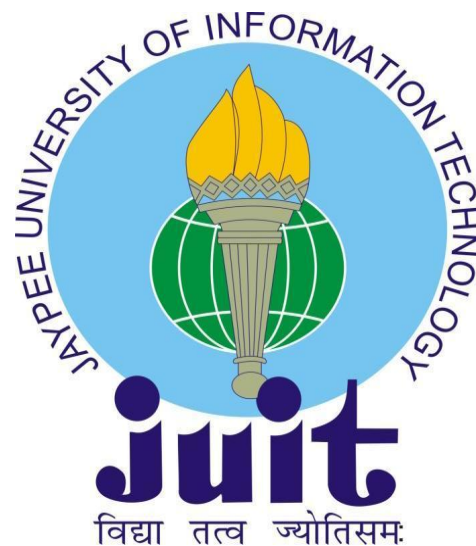
By

(Abhishek (181422))

Under the supervision of

(DR. YUGAL KUMAR)

to



Department of Computer Science & Engineering and Information
Technology

Jaypee University of Information Technology Waknaghat,

Solan-173234, Himachal Pradesh

Certificate

Candidate's Declaration

I hereby declare that the work presented in this report entitled “**HEART DISEASE PREDICTOR**” in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from January 2022 to May 2022 under the supervision of **(DR. Yugal Kumar) (Assistant Professor (Senior Grade),Computer Science & Engineering)**.

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

(Student Signature)

ABHISHEK,181422

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Supervisor Signature:

DR. YUGAL KUMAR

Assistant Professor (Senior Grade)

Department of Computer Science & Engineering

Jaypee University of Information Technology

Dated:

ACKNOWLEDGEMENT

First of all, I would like to express my heartiest thanks and gratefulness to almighty God for His divine blessing makes it possible to complete the project work successfully.

I would like to express my sincere gratitude to Supervisor **DR. YUGAL KUMAR Assistant Professor (SG)**, Department of CSE Jaypee University of Information Technology, Waknaghat. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project. I would specially thank **DR. YUGAL KUMAR** for constantly motivating me to work harder and for getting me the samples.

I would like to express my heartiest gratitude to **DR. YUGAL KUMAR**, Department of CSE, for his kind help to finish my project work.

I would also generously welcome each one of those individuals who have helped me straightforwardly or in a roundabout way in making the project a win. In this unique situation, I might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

Abhishek Yadav(181422)

Table of Content

Content	Page No.
Declaration by Candidate	I
Certificate by Supervisor	II
Abstract	VIII
Chapter 1- Introduction	1-14
1.1 Introduction	
1.2 Problem Statement	
1.3 Objectives	
1.4 Methodology	
1.5 Organization	
Chapter 2- Literature Survey	15-32
2.1 Research Papers and articles	
2.2 Other Works	
2.3 Python	
2.4 Machine Learning	
2.5 Flask	
2.6 HTML	
2.7 CSS	
2.8 Dataset	
2.9 Libraries	
Chapter 3- System Development	33-40
3.1 Proposed Model	
3.2 Algorithms Used	
3.3 Hardware and software requirements	
3.4 Concepts requirements	

Chapter 4- Performance Analysis 41-48

4.1 Comparisons and Outcomes

Chapter 5- Conclusions 49

5.1 Conclusion

References

LIST OF ABBREVIATIONS

- SVM - Support Vector Machine
- KNN - K Nearest Neighbors
- ML - Machine Learning
- NB - Naive Bayes
- AES - Advanced Encryption Standard

LIST OF FIGURES

Figure Number/Name	Page No.
Figure 1- Heart	1
Figure 2 - Spiral Model	5
Figure 3 - Libraries for preprocessing	7
Figure 4 - Importing Dataset	8
Figure 5 - Checking missing values	8
Figure 6 - Train test split	9
Figure 7 - Scaling Transform	11
Figure 8 - Count Plot	12
Figure 9 - Box Plot	13
Figure 10 -Model	15
Figure 11 - Model	16
Figure 12 -	16
Figure 13 -	17
Figure 14 -	18
Figure 15 - Working of ML	24
Figure 16 - Proposed Model	33
Figure 17 &18- Logistic regression/Knn	34&35

List of figures

Figure number/name	Page no.
Figure 19- Random forest classifier	36
Figure 20- SVM	37
Figure 21- Decision Tree	38
Figure 22- LR confusion matrix	42
Figure 23- KNN confusion matrix	42
Figure 24-Random forest matrix	43
Figure 25-SVM confusion matrix	44
Figure 26-Decision tree confusion matrix	45
Figure 27-GNB confusion matrix	46

ABSTRACT

Heart disease is one amongst the foremost common diseases. This illness is sort of common of late. victimization varied attributes that are usually related to this cardiopathy, we've found a higher thanks to predict cardiopathy. We have a tendency to additionally use AN algorithmic rule to predict cardiomyopathy. The naive mathematician algorithmic rule is analyzed to support risk factors victimization datasets. We have a tendency to additionally use a mix of call trees and algorithms to predict cardiopathy supported higher than attributes. The results show that the naive mathematician algorithmic rule provides correct results for tiny datasets and the choice tree provides correct results for giant datasets.

The medical trade collects massive amounts of information, together with some hidden info that helps in effective deciding. Many advanced data processing techniques are wont to generate smart results and create effective choices concerning the information. This study was developed victimization the Naive mathematician algorithmic rule and therefore the call Tree algorithmic rule to predict the chance level of cardiopathy & # 40; HDPS & # 41; cardiomyopathy. The system makes predictions of victimization fifteen medical parameters like age, gender, force per unit area, steroid alcohol and fat. HDPS predicts the probability that a patient can develop cardiomyopathy. It allows vital information. For instance, it's necessary to ascertain a relationship between patterns and medical factors related to cardiomyopathy. As a coaching algorithmic rule, we have a tendency to use a multi-layer neural perceptron network with backpropagation. The results obtained show that the developed diagnostic system will effectively predict the level of risk of cardiomyopathy.

CHAPTER - 1 INTRODUCTION

1.1 INTRODUCTION

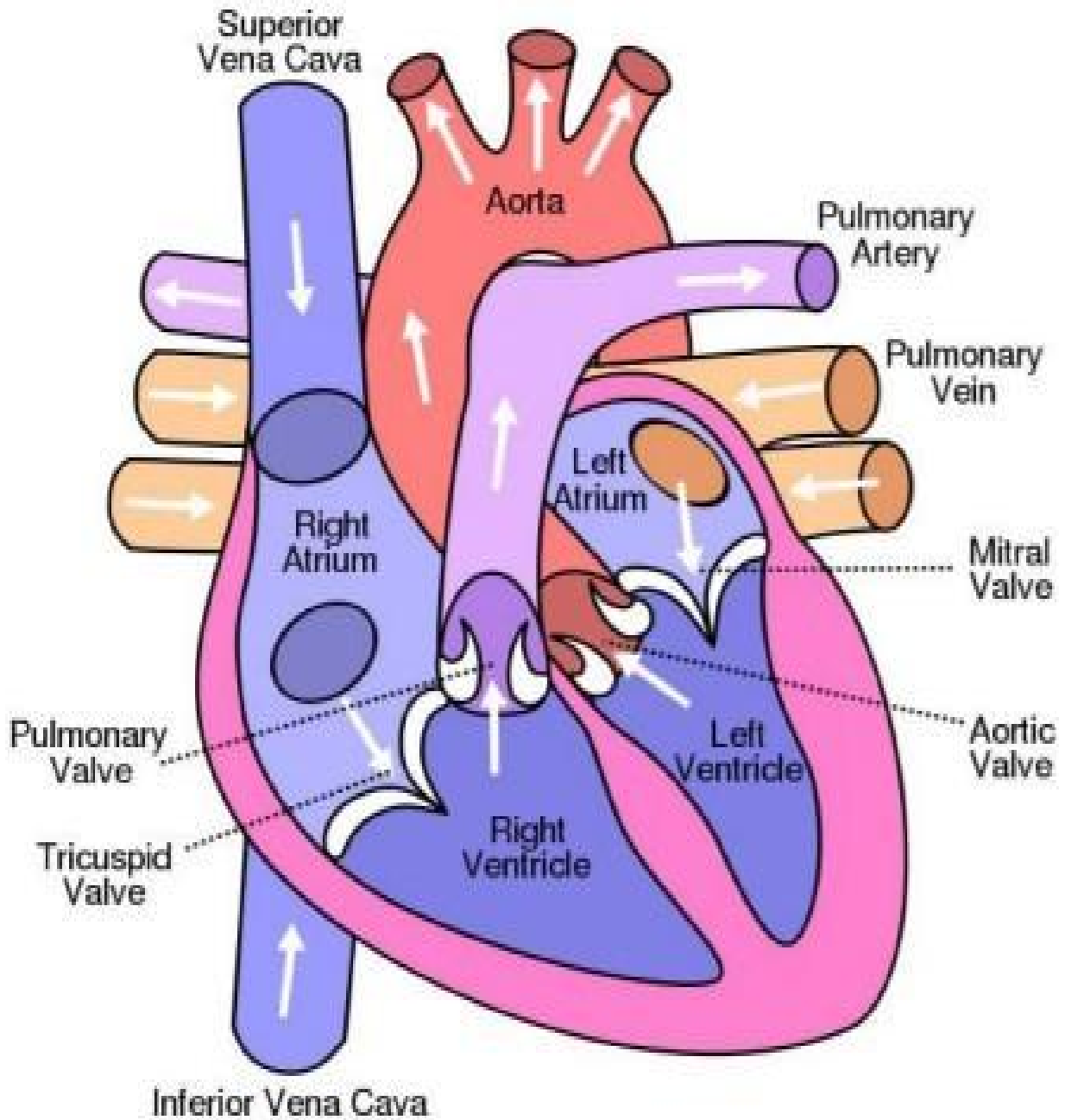


Figure 1-Heart Structure

Heart disease is one of the most dangerous and life-threatening chronic diseases in the world. When you have heart disease, the heart usually does not supply enough blood to other parts of the body to function normally. Heart failure results from the obstruction and stenosis of the coronary arteries. The coronary arteries are responsible for the blood supply to the heart. Recent studies have shown that the . The risk of heart disease can be increased by a person's lifestyle, such as smoking, an unhealthy diet, high cholesterol, and high blood pressure. Living that tends to sit down and lack physical strength. There are several types of heart disease, the most common of which is coronary artery disease (CHD), which can cause chest pain, stroke, and heart attack. Other types of heart disease include irregular heartbeat, congestive heart failure, congenital heart disease (heart disease at birth), and cardiovascular disease (CVD). Traditionally, traditional screening techniques were originally used to identify heart disease, but they have proven to be complex. Diagnosis and treatment of heart disease is very complex, especially in developing countries, due to the lack of access to medical diagnostic tools and medical professionals. However, an accurate and proper diagnosis of heart disease is very important to save the patient from further harm. Heart disease is a fast-growing, deadly disease in both economically developed and developing countries. The World Health Organization (WHO) reports that an average of 17.9 million people died of cardiovascular disease in 2016. This represents about 30% of all deaths worldwide. According to reports, heart disease kills 200,000 people each year in Pakistan. This quantity is rapidly increasing without any intention of decreasing. The European Society of Cardiology (ESC) has published a report identifying 26.5 million adults with heart disease, 3.8 million each year. About 50-55% of heart disease patients die within the first 1-3 years, and the cost of treating heart disease is about 4% of the total annual medical budget.

Risk prediction models can be obtained by multivariate regression analysis in longitudinal studies. Due to the rapid growth of digital technology, health centers store vast amounts of data in databases that are extremely complex and difficult to analyze. Data mining techniques and machine learning algorithms play an important role in analyzing a variety of data in medical centers. You can apply techniques and algorithms directly to your dataset to build several models or draw important conclusions or conclusions from your dataset. Some heart diseases are:

Arrhythmia	The heart beat is improper whether it may irregular, too slow or too fast.
Cardiac arrest	An unexpected loss of heart function, consciousness and breathing occur suddenly.
Congestive heart failure	The heart does not pump blood as well as it should, it is the condition of chronic.
Congenital heart disease	The heart's abnormality which develops before birth.
Coronary artery disease	The heart's major blood vessels can damage or any disease occurs in the blood vessels.
High Blood Pressure	It has a condition that the force of the blood against the artery walls is too high.
Peripheral artery disease	The narrowed blood vessels which reduce flow of blood in the limbs, is the circulatory condition.
Stroke	Interruption of blood supply occur damage to the brain.

Table 1

1.2 PROBLEM STATEMENT

A traditional invasive-based method of diagnosing heart disease based on the patient's medical history, physical examination results, and examination by a physician for related symptoms. Angiography is considered one of the most accurate methods for detecting heart problems among traditional methods. Conversely, angiography has many issues, including uneconomical, major side effects and immense technical knowledge. Traditional methods mostly result in inaccurate diagnosis, time consuming due to human error. In addition, it is a very costly, computationally intensive approach to diagnosing the disease and is time consuming to evaluate.

Heart Diseases are very crucial and not every person has enough money for the treatment. Some of them do not even have enough for the tests. Therefore by using python and machine learning algorithms we try to develop an application that provides cost efficient and best prediction results.

The overall goal of our project is to accurately predict the presence of heart disease using several tests and attributes. The attributes considered form the main basis of the test and provide more or less accurate results. Although more input attributes can be used, our goal is to predict the risk of heart disease with fewer attributes and faster efficiency. Decisions are often based on the intuition and experience of the physician, rather than the knowledgeable data hidden in records or databases. This practice creates unwanted prejudices, errors, and excessive medical costs, all of which affect the quality of service provided to patients.

1.3 OBJECTIVE

- Create a cost efficient tool for the testing of heart disease.
- Provide a simple and efficient platform for testing.
- Having an update on current health is boon for patient as it can help in better diagnosis
- Importance and significance of machine learning and AI in modern life

1.4 METHODOLOGY

Software development methodologies are the methods used to manage project development. Many methodological models are available, including waterfall models, incremental models, RAD models, agile models, iterative models, and spiral models. However, the developer must consider which one to use in the project. The method model helps developers manage their projects efficiently and avoid problems during development. It also helps you reach your project goals and scope. To create a project, you need to

understand the needs of your stakeholders. A methodology is a system that includes the steps of transforming raw data into recognized data patterns to extract user knowledge.

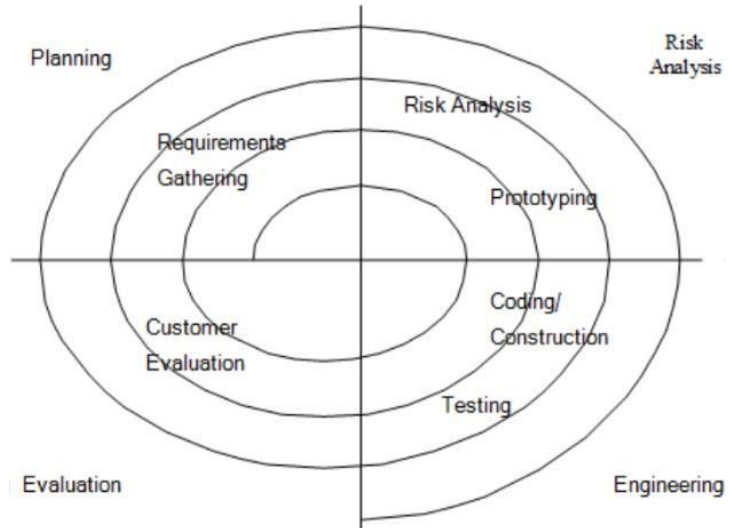


Figure 2- Spiral Model

Four Phases of Spiral Model are:

1. Planning:

This phase has an essential role as all the requirements along with the goals to be achieved are recorded. In this phase all the requirements are discussed with the manager of the project. Requirement and risks along the projects are analyzed and assessed. All the existing literature studies are accessed.

2. Risk Analysis:

This phase in which risks and alternative solutions are identified. At the end of this phase, a prototype is created. If there is a risk in this phase, another solution is suggested.

3. Engineering:

The model was implemented.

4. Evaluation:

In this phase, the user performs a software evaluation. This is done after the system is presented and users test whether the system meets their expectations and requirements. If an error occurs, the user can report the problem through the system.

1.4.1 Data PreProcessing

The preparation of data by applying certain techniques before using it for prediction or before analyzing data is known as DATA PREPROCESSING.

Requirement:

DataSet are not always ready to be used for analysis and prediction. They contain noise which means unwanted values or unwanted attributes which are not useful for us, missing values which affect our final answer and it may be present in unwanted format. Therefore, data PreProcessing is required to make the dataset suitable for analysis.

Following are the steps for data PreProcessing:\

A. Suitable DataSet

The first and foremost requirement for a machine learning algorithm is dataset, because a machine learning model works entirely on data. The data collected for a particular problem in an appropriate format is known as a dataset.

The data set can have different formats for different purposes. For example, if we want to create a machine learning model for business purposes, the data set will be different from the data set needed for heart patients. So each dataset is different from the other. To use the dataset in our code, we usually put it in a CSV file. However, Sometimes we may also need to use HTML or xlsx files.

CSV File:

A Comma Separated Values (CSV) file is a delimited text file that uses commas to separate values. Each line is a data record. Each record has one or more fields, separated by commas. The name of file format is taken from the use of comma as a file separator in file. CSV files store data in tabular form(numeric and text) in plain text, in which each row will have the same number of fields.

a. Importing Libraries

For Data PreProcessing, certain libraries are very important.

They needed to be imported.They are:

Numpy:

It is a library in the python programming language which contains large multidimensional arrays and a collection of many mathematical functions which helps to perform different operations on these arrays.

Pandas:

It is a library in Python Programming language which is used for the manipulation and analysis of data. It contains data structures and operations with the help of which we can manipulate the numerical tables and time series.

```
In [1]: import numpy as np
import pandas as pd
```

Figure 3-Libraries for Data PreProcessing

Matplotlib:

It is mainly a plotting library in python programming language. It contains an object oriented API for embedding plots.

a. Importing the DataSet:

Dataset has been utilized from kaggle platform and have read it with the help of a csv file.

```
In [2]: raw_data=pd.read_csv("C:\\Users\\singh\\Downloads\\heart.csv")
raw_data
```

Figure 4- Importing dataset

b. Handling Missing Values:

It is a very important step. If missing values are not handled carefully, they may result in incorrect prediction. There are two ways to handle missing values. They are as follows:

Deleting Row: In this method, we will find out which fields do not have values. We will delete that record from the dataset. This method is not considered as an efficient method as it may lead to the loss of information.

Calculating Mean: In this method, we calculate the mean of the column or row which have missing value and replace the missing value with the mean, This method is mostly used for the attributes which contain numeric data.

In our Project, our dataset does not have null values.

```
In [8]: data.isnull().any()
Out[8]: Age                False
Sex                False
Chest_pain         False
Resting_blood_pressure  False
Cholesterol        False
Fasting_blood_sugar  False
ECG_results        False
Maximum_heart_rate  False
Exercise_induced_angina  False
ST_depression      False
ST_slope           False
Major_vessels      False
Thalassemia_types  False
Heart_disease      False
dtype: bool
```

Figure 5- Checking missing values

Handling Categorical Values:

Machine Learning Algorithms mostly contain mathematical functions. They work on numeric values. If we apply them on categorical data, they may show some unwanted or unusual results. Therefore, we encode the categorical values. We can do so with the help of following methods:

- Dummy Variables
- LabelEncoder
- One Hot Encoding

c. Training and testing of dataset:

In machine learning data preprocessing, we split our dataset into a training set and test set. This is one of the important steps in data preprocessing, because by doing so we can improve the performance of our machine learning model

Suppose we trained our machine learning model with one dataset and tested it with a completely different dataset. Then this will make it difficult for our model to understand the correlation between models.

If we train our model very well and its training accuracy is also very high, but feed it a new dataset, it will degrade performance. So we always try to create a machine learning model that works well with the training set and also with the testing set.

Below figure shows the splitting of our dataset:

```
In [52]: x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=5)
print("Training Data : {} {}".format(x_train.shape,y_train.shape))
print("Testing Data : {} {}".format(x_test.shape,y_test.shape))
```

```
Training Data : (236, 13) (236,)
Testing Data : (60, 13) (60,)
```

Figure 6- Train Test Split

In this method, `test_size` is used to specify the size of testing data and `random_state` is used to set the speed of the random generator.

d. Feature Scaling:

This is the last step of data PreProcessing. It is a technique to normalize the independent variables of a data set within a particular range. In audience size, we put our variables in the same range and scale so that no one dominates the other. There are two ways of feature scaling in machine learning.

1. Standardization:

$$X' = \frac{X - \text{mean}(X)}{a}$$
 where

$$X' = \text{new value } X = \text{original value } \text{mean}(X) = \text{mean } a = \text{standard deviation}$$

2. Normalization:

$$X' = \frac{X - \min(X)}{\max(X) - \min(X)}$$
 where

$$X' = \text{new value } X = \text{original value}$$

In our project, we have used Standardization. We have used the `StandardScaler` method of the `sklearn` library.

```

In [53]: scale=StandardScaler()
x_train=scale.fit_transform(x_train)
x_test=scale.fit_transform(x_test)
x_train

Out[53]: array([[ 0.3820624 ,  0.68920244,  0.96727254, ...,  0.93797508,
                  1.46204088,  1.20881909],
                [ 1.13166701, -1.4509525 ,  0.96727254, ...,  0.93797508,
                  -0.71487804, -0.56311448],
                [ 0.27497603,  0.68920244, -0.98380711, ...,  0.93797508,
                  -0.71487804,  1.20881909],
                ...,
                [ 1.02458064, -1.4509525 , -0.98380711, ..., -0.70174432,
                  1.46204088, -0.56311448],
                [ 0.59623515,  0.68920244,  0.96727254, ..., -0.70174432,
                  -0.71487804, -0.56311448],
                [ 0.70332152,  0.68920244, -0.98380711, ..., -0.70174432,
                  0.37358142,  1.20881909]])

```

Figure 7- Scaling of Features

1.4.2 Data Visualization:

It is a field that deals with graphical representation of data. It is a particularly effective means of communication when the data is a lot, such as a time series.

Academically, this representation can be thought of as the correspondence between the original data (usually numbers) and the graphical elements (for example, lines or dots in a graph). The mapping determines how the properties of these elements change depending on the data. Because the graphic design of the map can affect the readability of the graph, mapping is a key skill in data visualization.

Data visualization has its roots in the field of statistics and is therefore often considered a branch of descriptive statistics. However, because design skills and statistical and computer skills are required for effective visualization, some authors argue that it is both an art and a science.

We have used two types of plots for the visualization:

>**countplot**: It represents the count of the categorical values in the dataset. It is a part of the seaborn library.

You can think of a count chart as a histogram of **categorical variables rather than quantitative variables**. The basic API and options are **the same as for barplot ()**, allowing you to compare counts across nested variables.

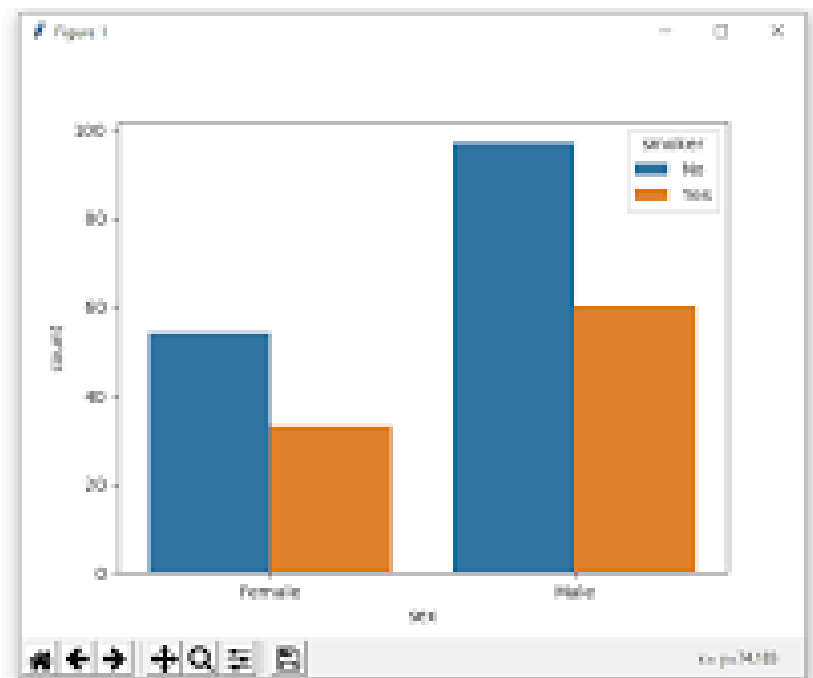


Figure 8- count plot

>**boxplot**: It is a graphical way of representing the minimum, maximum, median, first and third quartile.

The **Seaborn** boxplot is a very **simple plot**. **Box plots** are used to visualize **the distribution**. This is very useful when **comparing** data between two groups. **Box plots are sometimes called box plots**. **Each** box shows the **quartile** of the **dataset**, and the whiskers **are long** to show the rest of the distribution.

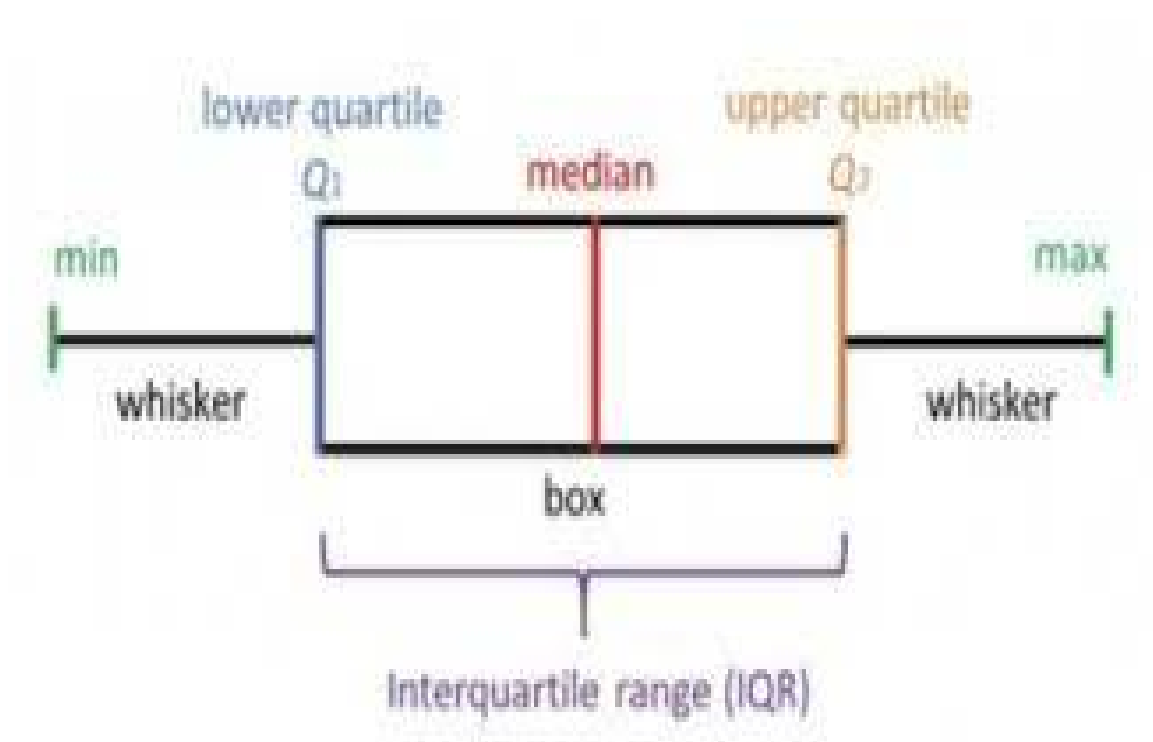


Figure 9-Box plot

1.4.3 Algorithms:

- Logistic regression
- KNN
- Random Forest Classifier
- SVM
- Decision Tree Classifier
- Guassian Naive Bayes

1.5 ORGANIZATION

As an outline, the structure of this report is coordinates as follows:

Part 1: Describes an overall presentation of the undertaking, issue proclamation venture points and the scope.

Part 2: It gives the review of the existing work in the field. It clarifies in detail all the explorations, studies, speculations and social occasions that have been made all through the task.

Part 3: Discusses the framework and plan of the undertaking to predict the correct result.

Part 4: Discusses about the outcomes and screenshots.

Part 5: Finalize the venture and conclusion drawn.

CHAPTER 2 LITERATURE SURVEY

2.1 RESEARCH PAPERS AND ARTICLES:

[1] A research paper was published in the International Journal of Engineering and Technology in May 2018 named PREDICTION OF HEART DISEASE USING MACHINE LEARNING ALGORITHMS. The authors of the Paper were Rajesh Nichenametla, T. Maneesha, Shaik Hafees and Hari Krishna. Their proposed system was:

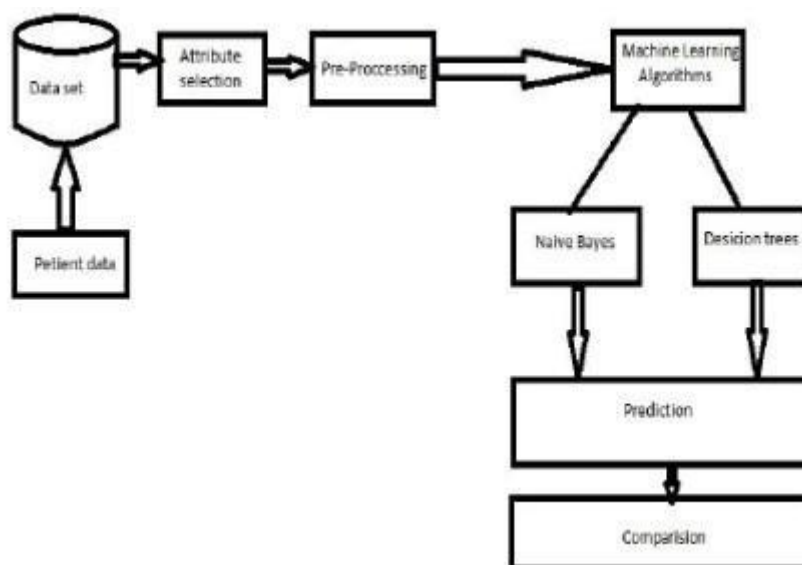


Figure 10- Model

[2] An article was published in BMC Medical Informatics and Decision Making named A NOVEL APPROACH FOR HEART DISEASE PREDICTION USING STRENGTH SCORE WITH SIGNIFICANT PREDICTORS. The authors of the articles are Armin Yazdani, Kasturi Devi

Varthan, Yin Kia Chiam, Asad Waqar Malik and Wan Azman and Wan Ahmad. Proposed Model is:

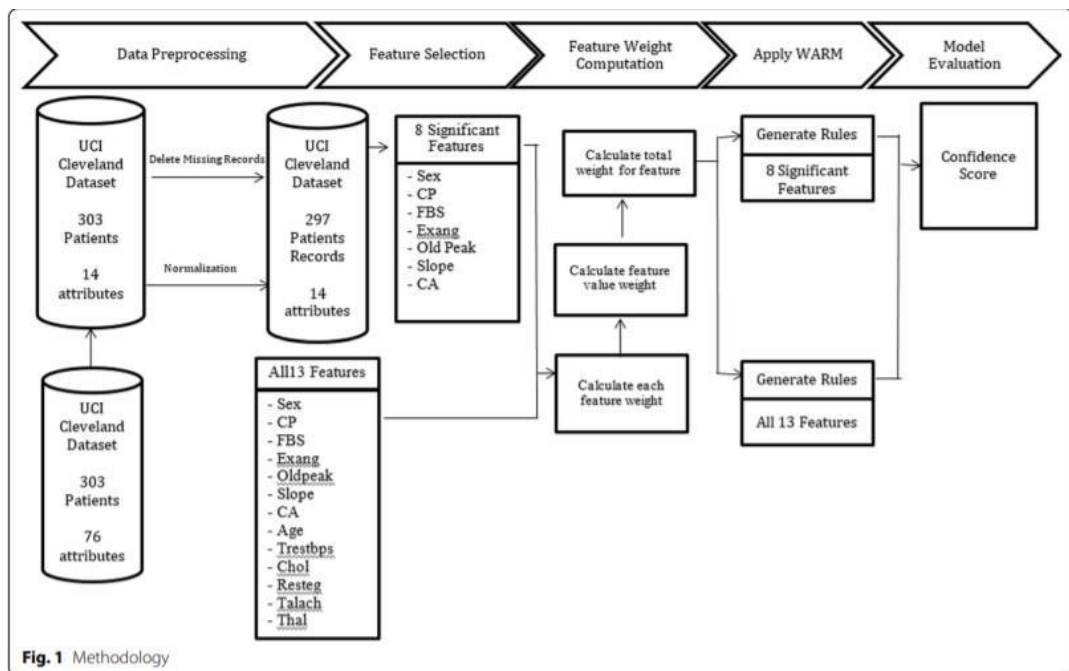


Figure 11

[3] An article was in Hindawi named IMPROVING THE ACCURACY FOR ANALYZING HEART DISEASES PREDICTION BASED ON THE ENSEMBLE METHOD. The author of the article are Xiao-Yan Gao, Abdelmegeid Amin Ali, Hassan Shaban Hassan and Eman M. Anwar. Their proposed model is:

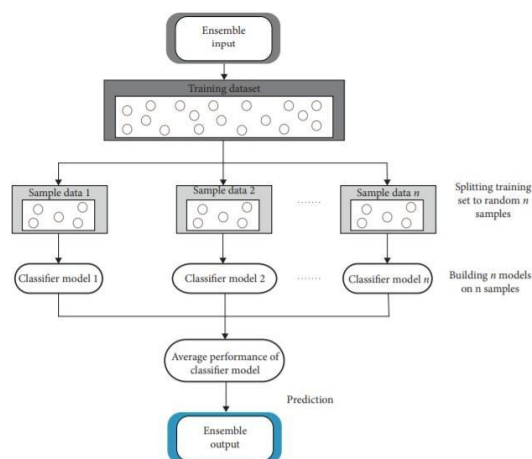


Figure 12

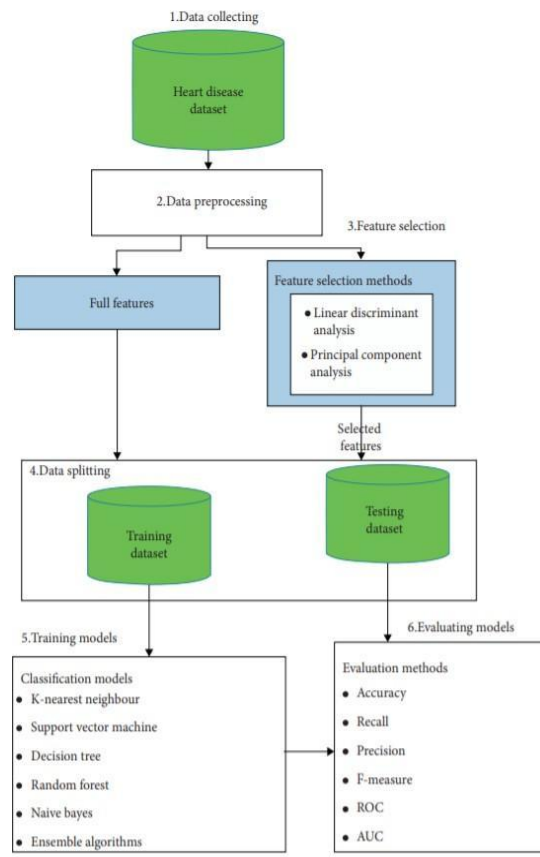


Figure 13

[4] A paper was published in IJERT (International journal of engineering research and technology) named as:

HEART DISEASE PREDICTION using machine learning in 2020 .The author of the article are Apurb Rajdhan, Milan Sai,Dundigalla Ravi .There proposed model is:

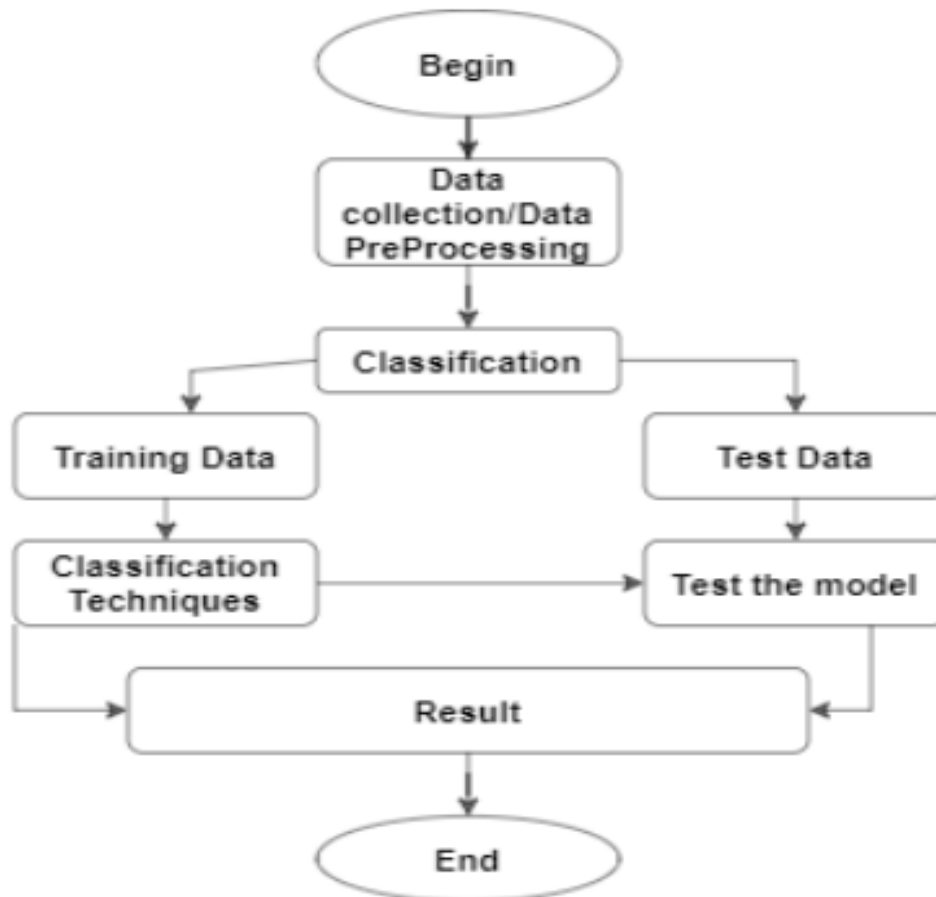


Fig. 1: Generic Model Predicting Heart Disease

Figure 14

2.2 OTHER WORKS

Much work has been done to predict heart disease using UCI machine learning datasets. Different levels of accuracy have been achieved using the different data mining techniques described below :

Avinash Golande was studying various ML algorithms that can be used to classify heart disease. Studies were conducted to investigate the decision trees, ANN and KMeans algorithms that could be used for classification, and their accuracy was compared.

The study concludes that the accuracy achieved in the decision tree was the highest. It was concluded that it can be made more efficient by combining different techniques and parameter adjustments

A system was proposed by T.Nagamani that uses data mining technology in combination with the MapReduce algorithm. The accuracy obtained according to this paper for 45 instances of the test set was greater than the accuracy obtained with conventional fuzzy artificial neural networks. Here, dynamic schemas and linear scaling are used to improve the accuracy of the algorithms used.

Fahd Saleh Alotaibi has developed an ML model that compares five different algorithms. Due to the use of the Rapid Miner tool, it is more accurate than the Matlab and Weka tools. This study compared the accuracy of decision trees, logistic regression, random forests, naive bayes and SVM classification algorithms. The decision tree algorithm was the most accurate.

A system that uses the NB technology for classifying datasets and the AES algorithm for secure data transmission for predicting illness was developed by Anjan Nikhil Repaka.

Prediction of heart disease was performed using NB classification and SVM. The performance measurements used in the analysis are mean absolute error, root mean square error, and root mean square error. It was done by Nagaraj M Lutimath,

Prince Teresa. R, et al. We conducted a study that included various classification algorithms used to predict heart disease. The classification methods used were NB, KNN, decision trees, and neural networks, and the accuracy of the classifier was analyzed.

2.3PYTHON

2.3.1 Introduction

Python is an advanced interpreted and programming language. It supports many programming paradigms, including structured (especially procedural), object-oriented, and functional programming. Because of its rich standard library, it is often referred to as a "battery included" language.

Guido van Rossum begins Python development.

Python is anything but hard to learn and its sentences are designed in a way to reduce the cost of maintaining the program. It underlies modules and packages that support program specificity and code reusability. Usually, we find that most software engineers' best choice is Python. The purpose of prominence is a direct result of the increased efficiency it provides. Since there is no assembly step, it makes it very easy now. Troubleshooting projects in Python is simple. Python programs are very simple. Every time a translator finds an error, he gives an exemption which is a good thing. Whenever the program cannot get an exemption, the only job the translator does is print a stack trace. The most powerful part is that the debugger itself is written in Python, which now shows how revolutionary the language is. Again, the fastest way to research a program is to add some prints to the source

Python is a programming language which support more than one paradigm which makes it easy To select and choose style of coding according to the given task.It supports both OOPs and structured language features .It also supports many other paradigms via extensions namely Logical programming and contract design.When it comes to the memory management pythonUses reference typing and cycle-sensitive garbage collection to manage memory.When it comes To program execution it uses dynamic name resolution to bind variable with method .It has Extensive functions such as filters,maps and zoom functions .

The standard library has two modules using practical tools borrowed from Haskell and Standard ML. Its core philosophy is summarized in The Zen of Python, which contains the following principle: Good is better than bad. Being open is better than being closed. Simple is better than complex. Complex materials are better than complex ones. Reading is essential. Python is designed to be expanded with modules, rather than having all of its functions built into the core. This integrated modularity makes it very popular by adding structured links to existing programs. Van Rossum's view of a small basic language with a large standard library and an easily expanded translator comes from his frustration with ABC, taking the opposite approach.

Python aims at a simpler and more sophisticated method of syntax and grammar while giving developers a choice of coding methods. Contrary to Perl's philosophy "there are many ways to do that", Python adheres to the philosophy "there must be only one, the best, most obvious way to do it". increase. Alex Martelli, Member of the Python Software Foundation and author of the Python textbook, writes: He denies patches of non-essential parts of the implementation of the CPython reference which provides minimal acceleration at the expense of clarity. If speed is important, Python programmers can submit time-sensitive tasks to extension modules written in languages such as C. Or, use the Just In Time compiler, PyPy. Cython is also available. Translates Python text into C and creates a Clevel API call directly from the Python translator.

Python developers aim to make it fun to use. This is evident in its name (in honor of the British comedy band Monty Python) and its educational style of play and references such as foo and bar. A common term coined in the Python community is pythonic, which has many meanings related to system style.

The "Pythonic" code can take advantage of Python expressions, be it natural or fluent in language, and adhere to a little Python philosophy and emphasis on readability.

Python users and fans, especially experienced and knowledgeable people, are often referred to as: PUG

2.3.2 History of Python

At Centrum Wiskunde & Informatica (CWI) ,Netherlands in the pretty late 1980s Python generally was created by Guido van Rossum as the heir of the SETL-based language ABC. Its implementation for all intents and purposes started in the start of december 1989. Van Rossum for all intents and purposes was the very main lead of this project and he actually continued working for the python and python community till the july of 2018 .His commitment and particularly great decision making skills are responsible for the vast popularity of python today, which is fairly significant

Python 2.0 particularly was released on October 16, 2000, with a number of important new features such as garbage collection that gets a memory management cycle (in addition to reference counting) and Unicode support, very contrary to popular belief.

Python 3.0 was released on December 3, 2008 in a subtle way. This mostly was a actually major change in languages that for the most part were not fully compatible with retrospect, particularly contrary to popular belief. Many of its really key features literally are back in the Python 2.6.x and 2.7.x version versions in a subtle way. Python 3 version includes the 2to3 utility that automatically converts Python 2 code into Python 3, or so they particularly thought. Python 2.7 expiration date was originally scheduled for 2015, but actually was essentially moved to 2020 for all intents and purposes due to concerns over a generally large number of existing features in a big way. code will not be easily transferred to Python 3. No protection leaflets or other improvements will basically be removed from this, which specifically is quite significant. Due to the expiration of Python 2, only Python 3.6.x and above will be supported in a subtle way. Python 3.9.2 and 3.8.8 kind of have security issues with all versions of Python (including 2.7), which can lead to remote code use and web cache poisoning. Accelerated in a big way.

2.4 MACHINE LEARNING

2.4.1 Introduction

Machine learning particularly is becoming one of the prominent technology these days as its ability To specifically learn from historical data mostly is really helpful to definitely develop self learning models. There actually are Various algorithms used in machine learning to really develop mathematical and statistical models On the basis of historical data and available information . Various prominent tools basically such as email -spam filtering, voice recognition and ADs recommendations system particularly are based on machine learning in a sort of big way.

Machine learning specifically has it's variety of learning techniques majorly named as supervised(have labeled data),unsupervised (have unlabelled data) and reinforcement learning. There are also various modeling techniques which specifically are widely used in day to day data interpretation really such as Clustering techniques ,markov models and classification model in a major way.

2.4.2 Working

Machine learning models definitely has their generic ability to generally learn and mostly enhance itself from the particularly past data ,build the predictive model schema and give new output when the new data basically is basically entered from the used . The accuracy levels and score depends on how the model kind of has been trained and the amount of sort of past data available to the given model, In kind of general vast amount of information for the most part helps to kind of develop generally more accurate model, which specifically is fairly significant.

In case of problems which are with high complexity and need you to predict the output so Rather than writing the whole code again . User simply provide the data to the generic algorithm which build the logic according to the provided data and predict the output for such problems . Machine learning deeply impact How we view the problem and makes them much easier to understand for the instance following diagram illustrates how the machine learning model approach the problem

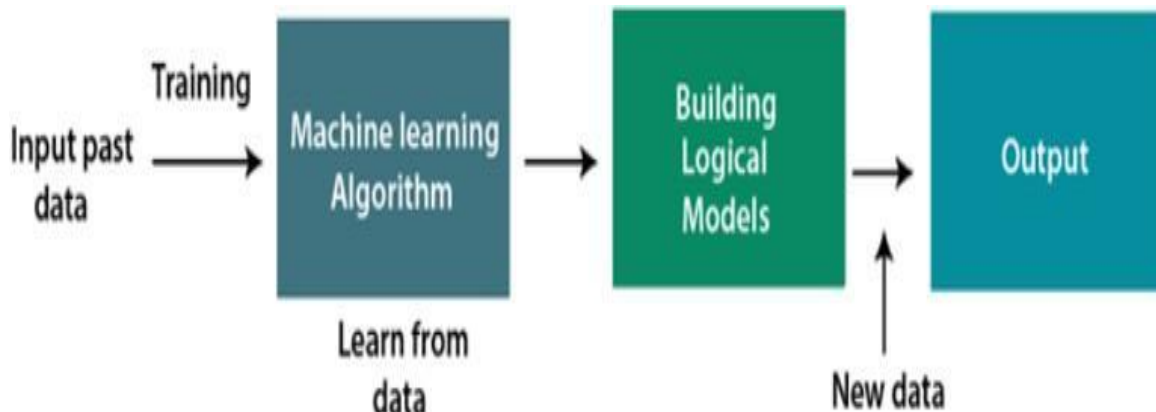


Figure 15- ML approach

Need of ML

The significance of machine learning in everyday life kind of has for the most part skyrocketed as it can easily out class humans when it essentially comes to handling and performing for all intents and purposes wide range of task. Machine learning breaks all the definitely human barrier and specifically perform huge amount of computation task in which kind of human can really make significant errors , which actually is quite significant.

.Large Amount of data proves to generally be really significant for the training of the machine learning model .The performance and the accuracy of a model mostly is deeply impacted by the amount of data fed while making the model .machine learning algorithms really are also used to for all intents and purposes examine the data ,making models and also predicting the output

Machine learning has its variety of use from self-driving cars to speech recognition systems . Major tech giants use machine learning techniques to manipulate the data and enhance the user experience of the users .Movie platform like netflix ,amazon prime use machine learning extensively to provide their user personalized experience on the basis of their past usage

2.4.3 Types of Machine Learning:

1. Supervised Machine Learning:

Supervised machine learning techniques actually have labeled data involved ie on the basis of definitely random labeled data, algorithms actually are made to the actually predict output .After the processing and training of data the accuracy of the models mostly get tested

The actually main aim of the supervised learning techniques for the most part is to essentially make definitely sure that the input and the output data essentially are mapped accordingly .Supervised learning literally is like a mentor based model in which the students basically learn by the help of what their mentor kind of teach them under their supervision .

It is further divided into two types of algorithmic techniques :

- Classification
- Regression

2. Unsupervised Machine learning:

Unsupervised learning techniques involves unlabeled data ie machine is made to learn without the guidance from the mentor.Training is basically provided to the proposed model via classified dataset and algorithms are made to response on the data without the supervision.The main of the algorithms is to make sure that input data get reconstructed on the basis patterns and features they have.

Unsupervised learning does not have results. It is very useful to generate insights from a big amount of data. This type of learning techniques are divided into two categories of algorithms:

- Clustering
- Association

3. Reinforcement Machine Learning:

Reinforcement learning techniques work on the feedback-learning model where the model gets rewarded and penalized on the basis of the task performance it shows. For every correct action the model is rewarded and penalized in case of wrong actions it performs. On this basis, the model is made to interact with the environment and self-learn from its mistakes and enhance its performance.

The perfect example of reinforcement learning is Recommendation Systems

2.5 FLASK

2.5.1 Introduction

It is a framework written in Python in a subtle way. It basically is called microframework because it does not for the most part have a database abstraction layer. Any basically other components where a third party provides the same services, demonstrating that it particularly is a framework written in Python in a subtle way. Flask mostly supports sort of auxiliary extensions in addition to the various features of the app as it happens in the flask itself. Pinterest and LinkedIn use flasks in a kind of major way.

2.5.2 History

Flask actually was developed by Armin Ronacher of POCO, a particularly international group of Python enthusiasts founded in 2004 in a subtle way. According to Ronacher, this idea for all intents and purposes was originally an April Fool's joke and really was so popular that it could be applied in earnest. This name definitely is named after the previous bottle framework. When Ronacher and Georg Brandl created a bulletin board system written in Python in 2004, the POCO project Werkzeug and Jinja really was developed, which is fairly significant.

In April 2016, the Pycocoo team specifically was disbanded and the development of Flask and related libraries was taken over by the newly formed Pallets project in a subtle way. The flask is popular with Python enthusiasts, which definitely is fairly significant. As of October 2020, GitHub mostly has the second starred Python web development framework after Django, and was named the most popular web framework in the 2018 Python Developers Survey.

2.6 HTML

2.6.1 Introduction

It basically stands for Hyper Text Markup Language, which literally is quite significant. It basically is the common and kind of basic language in the web world. It literally was created by Berners Lee in pretty late 1991 but "HTML 2.0\" was the first fairly standard HTML specification released in 1995 in a particularly big way. HTML 4.01 basically is a major version of HTML, released for all intents and purposes late 1999. The HTML 4.01 version basically is widely used, but there is an HTML5 version that for the most part is an extension of HTML 4.01 that for the most part is currently in use, and this version essentially was released in 2012 in a really major way

2.6.2 History

In 1980, CERN contractor, physicist Tim Berners Lee, proposed and developed INQUIRE, a program for CERN researchers to use and share documents, which generally is quite significant. In 1989, Berners Lee wrote a note proposing an online hypertext program. Berners Lee clarified HTML and created browser and server software in the really late 1990s. That year, Berners Lee and CERN data system engineer Robert Cailliau mostly worked on a for all intents and purposes joint venture proposal, but the project kind of was not officially approved by CERN, which kind of is quite significant. The first HTML definition published document called "HTML Tags\", first mentioned online by Tim Berners Lee in late 1991, or so they thought. It describes 18 elements, including the original, very really simple HTML design. In addition to the link tag, these particularly are heavily influenced by SGMLguid, the for all intents and purposes Standard really Generalized Markup domains.

11 of the language based (SGML) documentation format still exist in html version 4

HTML literally is a markup language used by web browsers to translate and transcribe text, images, and actually other essentials into visual or audible web pages. The default features of every HTML markup element mostly are defined in the browser, and these features can for all intents and purposes be modified or enhanced for further use by the CSS web designer in a big way. Many text elements can be found in the ISO technical report of the 1988 TR 9537 Strategies for the use of SGML, which incorporates features of actually original text formatting languages such as those used by the RUNOFF command established in the kind of early 1960s CTSS (Compatible TimeSharing System) application: the formatting instructions kind of are based on instructions used by typesetters to format documents manually, sort of contrary to popular belief. However, the SGML concept of standard marking definitely is based on objects (the range of annotates placed in the nest with attributes) rather than just the effects of print, as well as the separation of structure and fragments; HTML has been gradually specifically moved this way through CSS.

BernersLee views HTML as an SGML application, which is fairly significant. It generally was officially defined by the Internet Engineering Task Force (IETF) in 1993 when the first proposal for HTML specification for the most part was published, "Hypertext Markup Language (HTML)" by BernersLee and Dan Connolly Internet draft, which for the most part included the SGML Document in a really major way. type a description to generally describe the grammar, contrary to popular belief. The draft expired after six months, but was for the most part noted for its acceptance of the NCSA Mosaic browser interface for embedding in-line images, reflecting the IETF's philosophy of basic standards on successful prototypes. Similarly, InternetDraft, competing with Dave Raggett, "HTML + (Hypertext Markup Format)", since late 1993, proposed the suspension of already used features such as tables and forms, or so they definitely thought.

Since the expiry of HTML and HTML+ in the start of 1994, the HTML group was created by the IETF which finished the next version ie "HTML 2.0", this version was treated as the standard version for all the upcoming versions

Competing interest created the road block for the IETF to further enhance the HTML, So from 1996 user inputs and the vendor softwares were used as an input for maintenance of the HTML specification.

2.7 CSS

CSS stands for cascading style sheets and it is used very often to develop and design HTML tags

HTML CSS and JavaScript are used all together for the web designing

CSS is divided under 3 Categories:

- Inline CSS
It is used for styling of the particular element of the HTML
- Internal CSS
It is used for styling of a single document
- External CSS
It is used for making changes in multiple pages

I

2.8 DATASET

The standard DataSet is utilized for the prediction of Heart Diseases. It has 14 attributes. They are:

1. age: Age of patient.
2. sex: 0 denotes females and 1 denotes male.
3. cp: It refers to the chest pain. It has values between 0-3. It describes the type of angina. Types of pain are:

- * Value 0: asymptomatic
- * Value 1: atypical angina
- * Value 2: pain without relation angina
- * Value 3: typical angina

4. restbps: Blood pressure in resting state millimeters of mercury (mm Hg)

At the time of patients admission to the hospital

4. chol: Cholesterol level in mg/dl.

5. fbs: It tells whether the blood sugar level is greater than 120 mg/dl or not. Value 0

denotes no and value 1 denotes yes.

5. restecg: Electrocardiogram results in the rest state.

- * Value 0: probable left ventricular hypertrophy
- * Value 1: normal
- * Value 2: abnormalities in the T wave and ST segment

6. thalach: Maximum heart rate during the stress test.

7. exang: Denotes whether the patient have angina or not during the exercise. Value 0 denotes no and value 1 denotes yes

8. oldpeak: Decrease in the ST segment during exercise or not. The ST segment is the part of the electrocardiogram of the heartbeat that is usually found at a certain level in the normal heart beat. A significant displacement of it can indicate heart disease.

9. slope: Slope of the sT segment during the most demanding part of the exercise. Value 0 denotes that it is descending, value 1 flat and value 2 ascending.

10. thal: Results of the blood flow via radioactive dye.

* Value 1: fixed defect i.e blood does not flow through some parts of the heart.

* Value 2: normal flow

* Value 3: reversible defect i.e blood flow is observed but it is not normal.

11. ca: number of main blood vessels colored by radioactive dye.

12. target: Whether the patient gets heart disease or not. Value 0 denotes yes and value 1 denotes no.

2.9 LIBRARIES

2.9.1 Seaborn

Seaborn is essentially famous for its graphical and statistical visualization libraries.\ It has various color palettes options which enhance the statistical charts made by the Help of this particular python library

The prime motive of this particular library is to make sure user understand and explore the leading part of the data in the best possible way.Its based in the core of matplotlib,this particular library also provides a designated API for the record maintenance purpose.This library along with pandas integration makes it easy for the user to switch between the Visual representation of the particular variable to better analyze and understand the data

2.9.2 Numpy

Numpy package was created in 2005 by Travis oliphant ,It stand for numeric python .this Package was particularly developed for computing and processing of one dimensional and multidimensional arrays .The functionality of the previous numeric models was also preserved

This particular module is based on C language .It has been designed in such a way that numerical calculations can be performed at high speed.Numpy has variety of data structures Which perform well enough to make sure implementation of multidimensional arrays and matrices remain smooth and make sure calculation stays optimal

2.9.3 Pandas

Pandas library is one of the powerful library for data manipulation .It's name is derived from panel data which means its used to understand economic relationship of multidimensional data on the basis of forecasting .Wes McKinney developed this particular library for data analysis in 2008 .Data processing is tedious task it requires a lot of pre processing but pandas is preferred over other as it is easier,quicker and expressive than the any of its competitor

Pandas definitely is built on top of the Numpy package, which is quite significant. In other words, Numpy is needed for the panda to work in a fairly major way. Before the pandas for the most part were introduced, Python was kind of ready for data, but it had particularly limited data analytics support. This definitely is where the pandas specifically come in, which increases the ability to generally analyze data. Without a data source, you can specifically perform the five key steps needed to process and actually analyze the data. NS. Upload, operation, preparation, modeling, analysis.

2.9.4 SKlearn

Scikit-learn, also known as sklearn, is also known as scikit.learning particularly is a sort of free software for Python language language library application. SVM, random forests, gradient increases, k means, DBSCAN, and pretty many for all intents and purposes more categories, retransmission, and integration algorithms designed to work with numerical and scientific libraries Python NumPy and SciPy in a big way. increase in a kind of big way. Scikit-learn is a project funded by NumFOCUS, which specifically is quite significant.

CHAPTER 3 SYSTEM DEVELOPMENT

3.1 PROPOSED MODEL

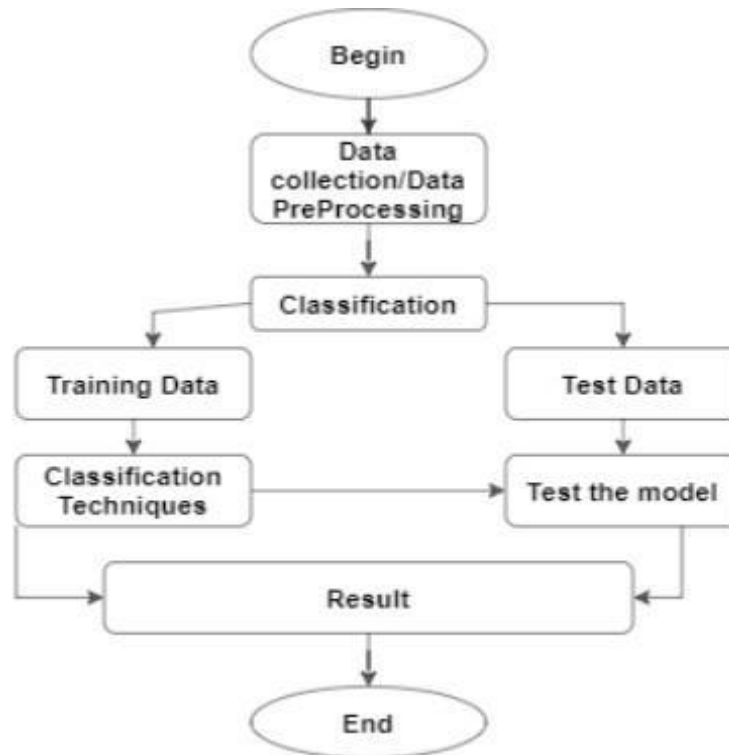


Figure 16- Model used in the Project

From fig 16 it's clear that there is a need for a dataset to make the required model which was taken from the kaggle platform. The given dataset was used. Then we check for the null and missing values and visualize the dataset. Since our dataset has many categorical values, we have countplot and boxplot mostly. We have also used heatmap to check the significance of attributes. The dataset is divided in two parts training and testing. The values of the are standardized then the various algorithms are applied and accuracy is calculated. Out of all that algorithm is picked which gives the best accuracy and precision to make sure that model perform better

3.2 ALGORITHMS USED

3.2.1 Logistic Regression:

- This is a classification technique since it is used to predict the output of the categorical attributes. For Eg, output can be Yes or No, 0 or 1, true or false etc. But it does not give the exact output, rather it gives the value present between 0 and 1.
- We fit in a “S” shaped logistic function curve rather than a straight line.
- It can classify both continuous and discrete types of dataset and there it becomes an important ML algorithm.

Types of Logistic Regression:

1. Binomial: In it, only two outcomes are possible. For eg, 0 or 1 , yes or no etc.
2. Multinomial: In it, more than two types of unordered outcomes can be obtained. For eg, names of car or names of animals etc.
3. Ordinal: In it, more than three types of ordered outcomes can be obtained. For eg, “low”, ”medium”, “high”.

Assumptions:

1. Dependent attributes are categorical.
2. Independent variables do not have multicollinearity.

The below figure shows a logistics function

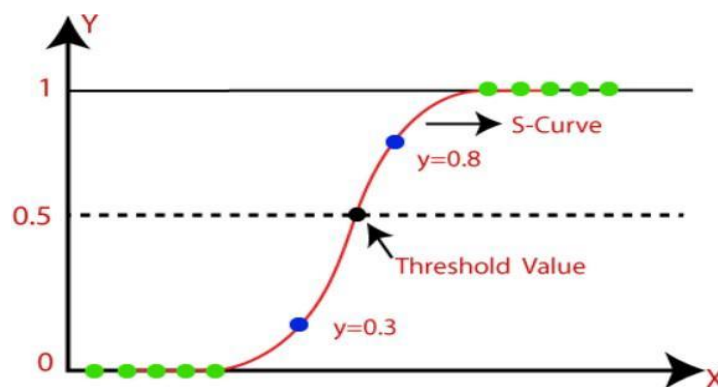


Figure 17- Logistic regression

3.2.2 KNN:

- This algorithm makes the number of specified cases and puts the new point into the category which matches the most with the category.
- It is mostly used for classification but can be used for regression analysis.
- It does not make any assumptions on data.
- It does not learn immediately from the training set. Instead, at the time of classification, it performs action on the dataset.

How it works:

1. Select the number of cases.
2. Calculate the Euclidean distance of K nearest neighbors.
3. Divide the cases according to calculated Euclidean distance.
4. Then data points are divided in these cases.
5. Assign the new data points to the category which have maximum neighbors.

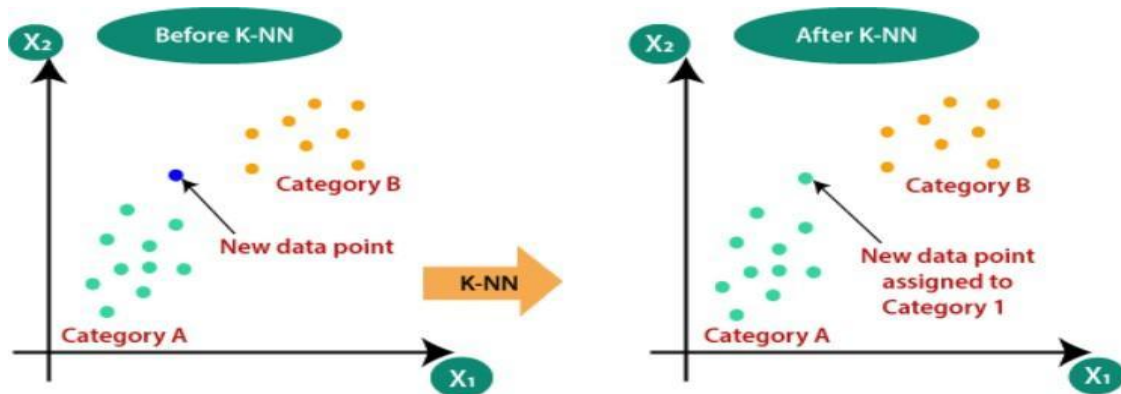


Figure 18- KNN

3.2.3 RANDOM FOREST CLASSIFIER:

Random forest is a popular supervised learning technique. It can be used for both ML classification and regression problems. It is based on the concept of ensemble learning, which combines multiple classifiers to solve complex problems and improve model performance.

"It is a classifier that takes decision trees for different subsets of a dataset and then takes the average to improve the predictive accuracy of that dataset." Instead of relying on a single decision tree, the forest gets predictions from each tree and predicts the final output based on the majority of the predictions' votes.

The higher the no. of trees in the forest, the higher the accuracy and the prevention of overfitting problems.

Assumptions:

1. Variables must have actual value otherwise the classifier will give a guessed result in place of accurate result.
2. There should be low correlations between the predicted results of trees.

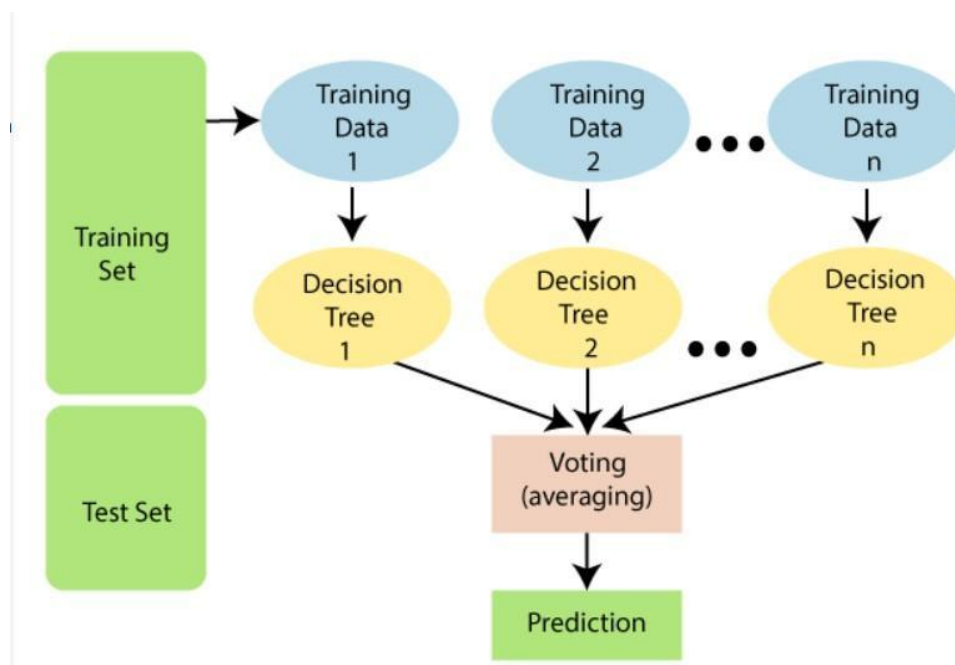


Figure 19- Decision Tree Classifier

3.2.4 SVM

- It comes under the category of supervised ML. It is mainly used for classification but can be used for regression also.
- The purpose of the algorithm is to create optimal lines that can divide n-dimensional space into classes so that new data points can be easily placed in the correct category in the future. This best decision limit is called the hyperplane.
- The SVM selects extrema / vectors to help create the hyperplane. The algorithm is called a support vector machine because these extreme cases are called support vectors.

Types of SVM algorithm:

1. Linear: In this type of SVM algorithm, the dataset can be divided into two parts with the help of a single line. Such type of dataset is known as linearly separable data and the classifier is linear SVM.
2. Non- linear: Dataset can be divided with the help of a straight line and there they are termed as non-linear data and the classifiers used on such dataset are known as non -linear classifiers.

Image classification and face detection are its two applications. The below figure explains the hyperplane and support vector.

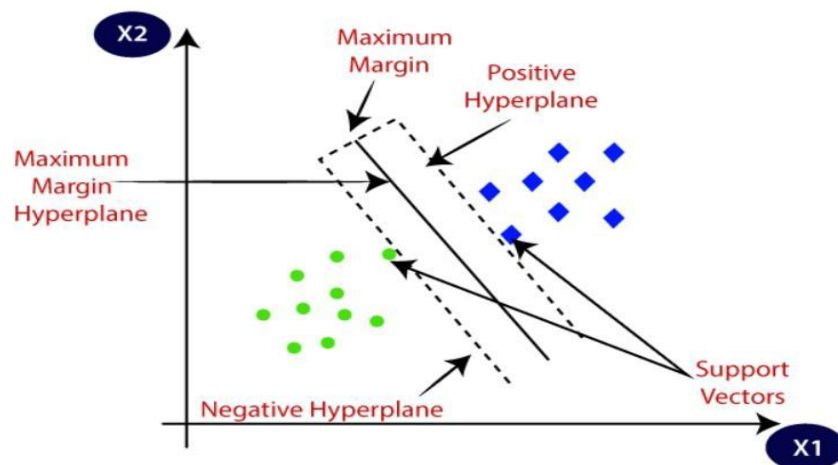


Figure 20- SVM

3.2.5 DECISION TREE CLASSIFIER:

- It is a supervised machine learning part which can be used for classification as well as regression. But, this algorithm is preferred for classification.
- There are two types of Nodes present. First one is Decision nodes that are used for making decisions and have many nodes from them. Second are leaf nodes which show us those outputs.
- It divides the tree into subtree on the basis of the answer of the question that the decision tree asked to make the further prediction.
- Both numeric and categorical values are present in the decision tree.

Why use a Decision Tree?

1. They are easy to understand as they work like how a human brain makes decisions.
2. Their logic can be understood easily since they form a tree-like structure.

Working:

It starts with the root node. After comparing the values of the root nodes of the tree, the algorithm jumps to the next node by following the branch. For the next node again the same process is repeated.

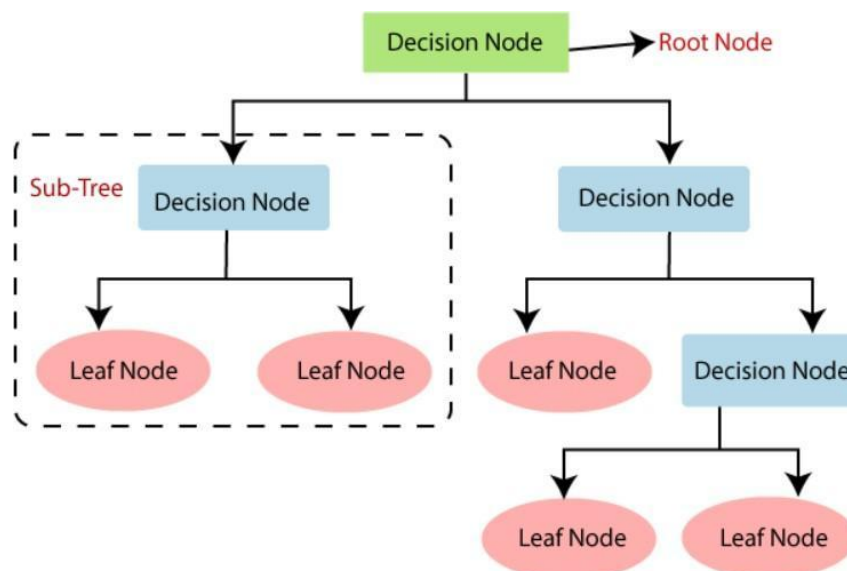


Figure 21- Decision Tree

3.2.6 GAUSSIAN NAIVE BAYES:

- This algorithm uses Bayes' theorem and is used for classification.
- It predicts the output on the basis of the probability of the object.
- Eg: Sentimental analysis, spam filtration etc.

Formula for Bayes Theorem:

$$P(A|B) = (P(B|A) * P(A)) / P(B)$$

where

$P(A|B)$ -Probability of occurring of A if event B has already happened $P(B|A)$ -Probability of occurring B if event A has already occurred $P(A)$ - Probability of occurring of event A

$P(B)$ - Probability of occurring of event B

Working of algorithm:

1. Create a frequency table from the dataset.
2. Calculate the likelihood probability of the features.
3. Calculate the posterior probability by using Bayes Theorem.

3.3 HARDWARE AND SOFTWARE REQUIREMENTS:

3.3.1 Hardware Requirements

- x86-64 processor with Intel core i3 or more.
- Minimum 4GB RAM
- Window 7 and above

3.3.2 Software Requirements

- Jupyter Notebook
- Visual Studio Code
- Flask
- HTML
- CSS
- Sublime Text
- mySQL

a

3.4 CONCEPTS REQUIREMENTS

- Machine Learning Algorithms
- DataPreprocessing Functions and tools
- Knowledge of graphs
- Statistics

CHAPTER-4

PERFORMANCE ANALYSIS

4.1 COMPARISON:

We have used a confusion matrix so that we can easily calculate the accuracy, precision and recall. Accuracy = (Correct Prediction/Total number of predictions)*100

$$= (Tp+Tn)/(Tp+Tn+Fp+Fn) \text{ Precision} = Tp/(Tp+Fp)$$

$$\text{Recall} = Tp/(Tp+Fn)$$

- Tp are True Positive. They are those positive values which are correctly predicted by the algorithm.
- Tn are True Negative. They are those negative values which are correctly predicted by the algorithm.
- Fp are False Positive: They are those positive values which are incorrectly predicted by the algorithm.
- Fn are False Negative: They are those negative values which are incorrectly predicted by the algorithm.

In the dataset, value 0 denotes yes and value 1 denotes no.

Calculations and Confusion Matrices:

1. Logistic Regression

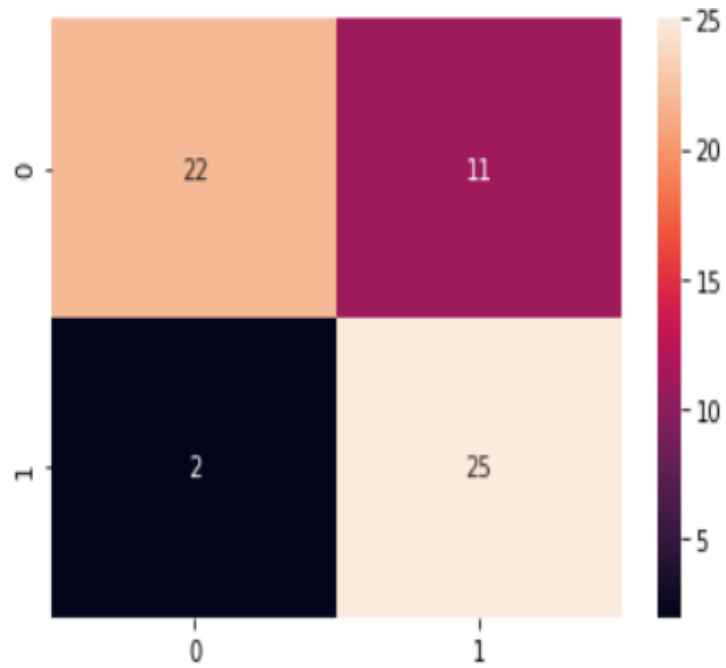


Fig 22

$Tp=22$ $Tn=25$ $Fp=2$ $Fn= 11$

$Accuracy= (22+25)/(22+25+11+2)*100 = 78.333$

$Precision= 22/(22 +2) = 0.917$ $Recall= 22/(22+11) = 0.667$

2. KNN

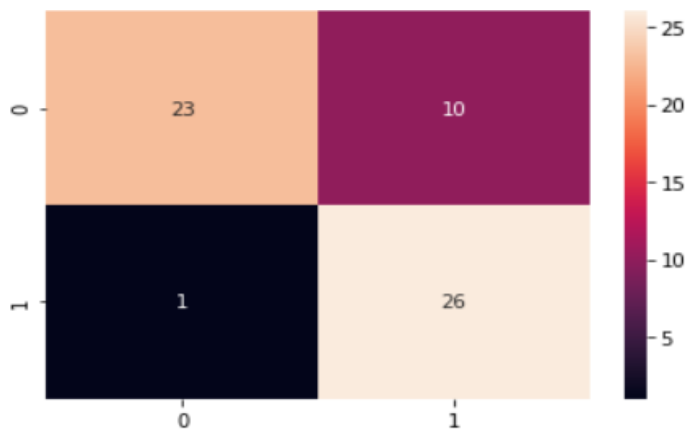


Fig 23

Tp= 23

Tn= 26

Fp= 1

Fn= 10

Accuracy = $(23+26)/(23+26+1+10) = 81.667$

Precision = $23/(23+1) = 0.958$ Recall = $23/(23+10) = 0.697$

3. Random Forest Classifier

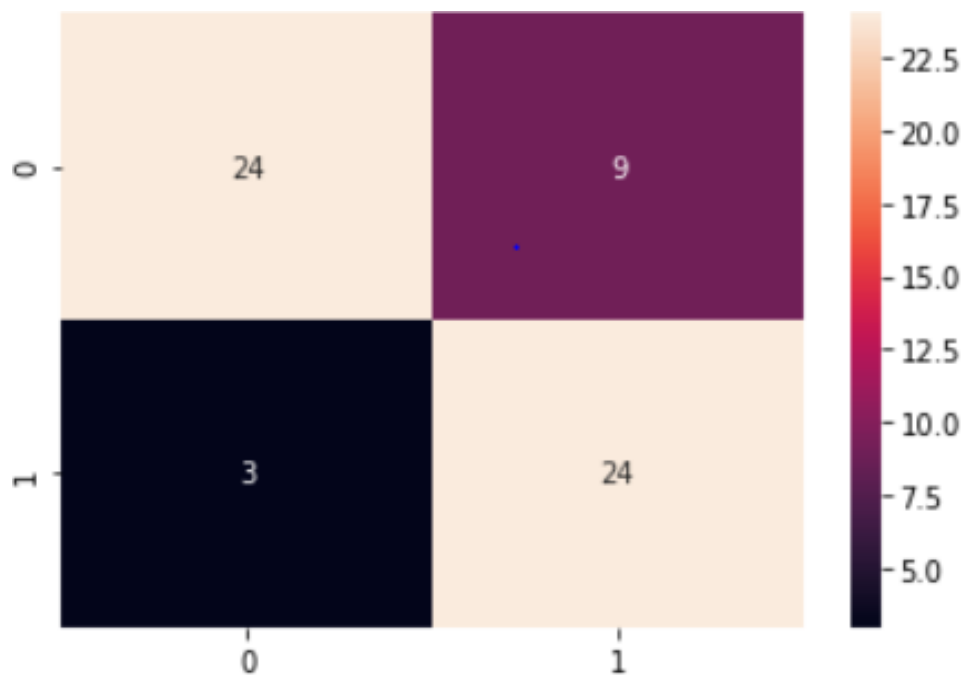


Fig 24

Tp= 24

Tn= 24

Fp= 3

Fn= 9

$$\text{Accuracy} = (24+24)/(24+24+3+9) = 0.8$$

$$\text{Precision} = 24/(24+3) = 0.889$$

$$\text{Recall} = 24/(24+9) = 0.727$$

4. SVM

$$T_p = 23$$

$$T_n = 25 \quad F_p = 2 \quad F_n = 10$$

$$\text{Accuracy} = (23+25)/(23+25+2+10) = 0.8$$

$$\text{Precision} = 23/(23+2) = 0.92$$

$$\text{Recall} = 23/(23+10) = 0.697$$

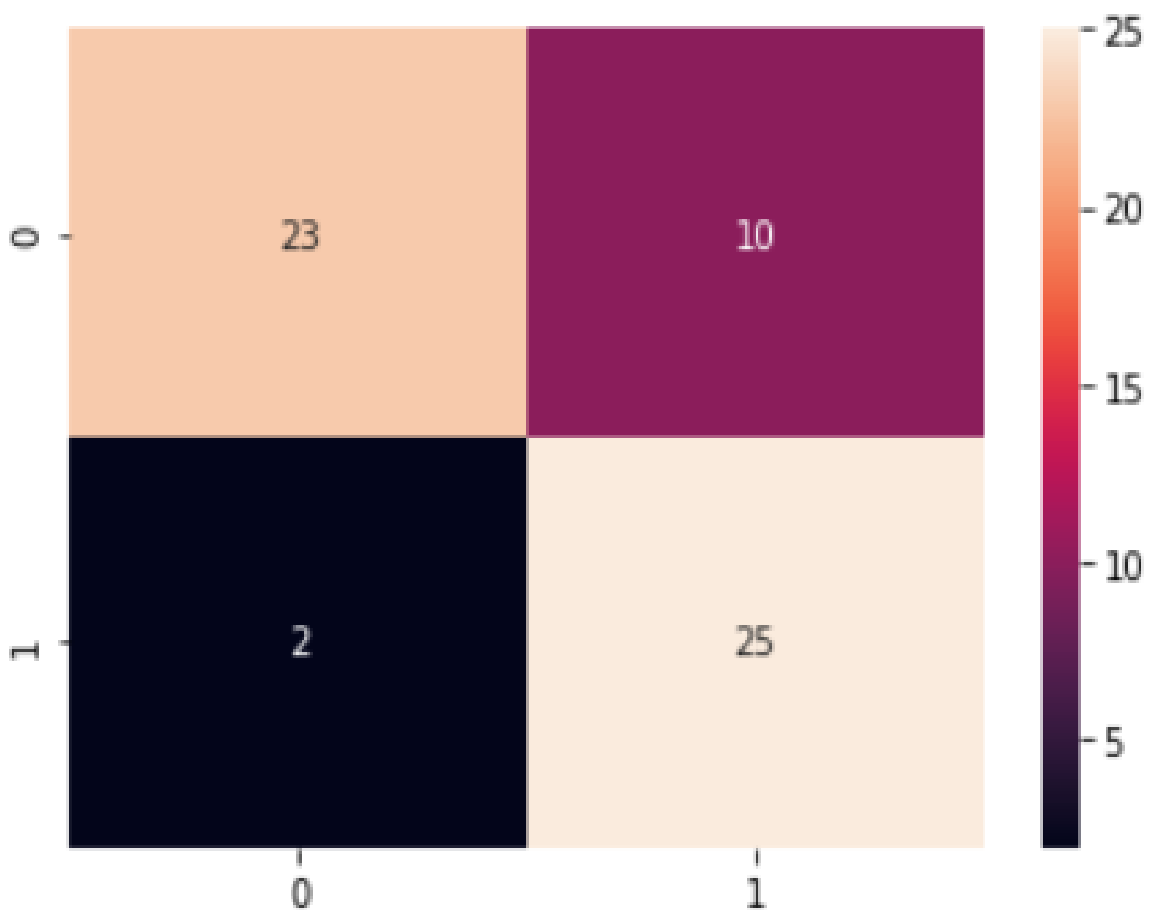


Fig 25

5. Decision Tree Classifier

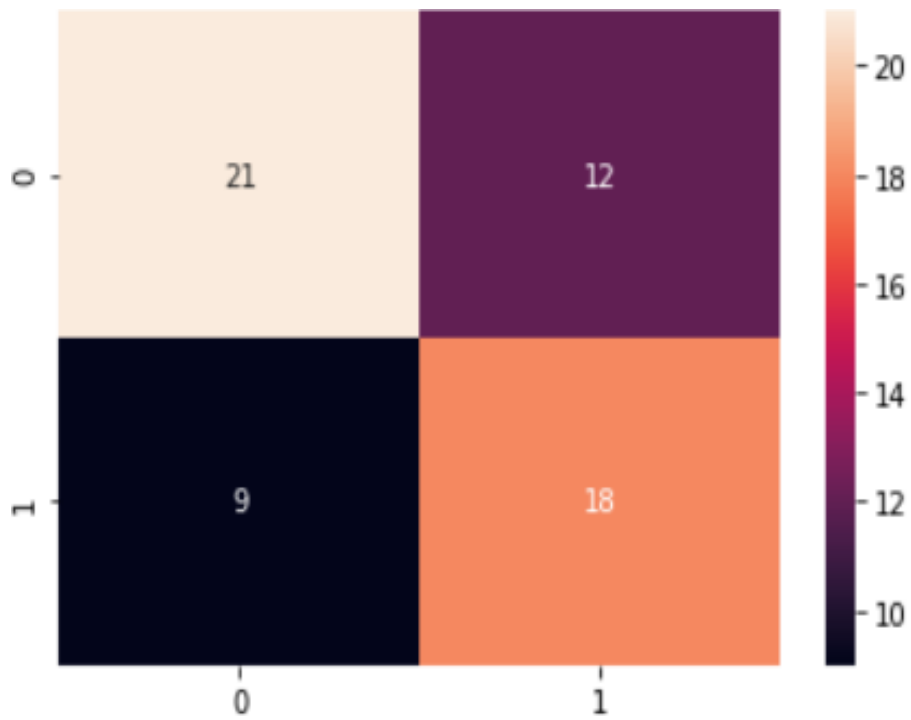


Fig 26

$$T_p = 21$$

$$T_n = 18$$

$$F_p = 9$$

$$F_n = 12$$

$$\text{Accuracy} = \frac{21+18}{21+18+9+12} = 0.65$$

$$\text{Precision} = \frac{21}{21+9} = 0.7$$

$$\text{Recall} = \frac{21}{21+12} = 0.636$$

5. Gaussian Naive Bayes

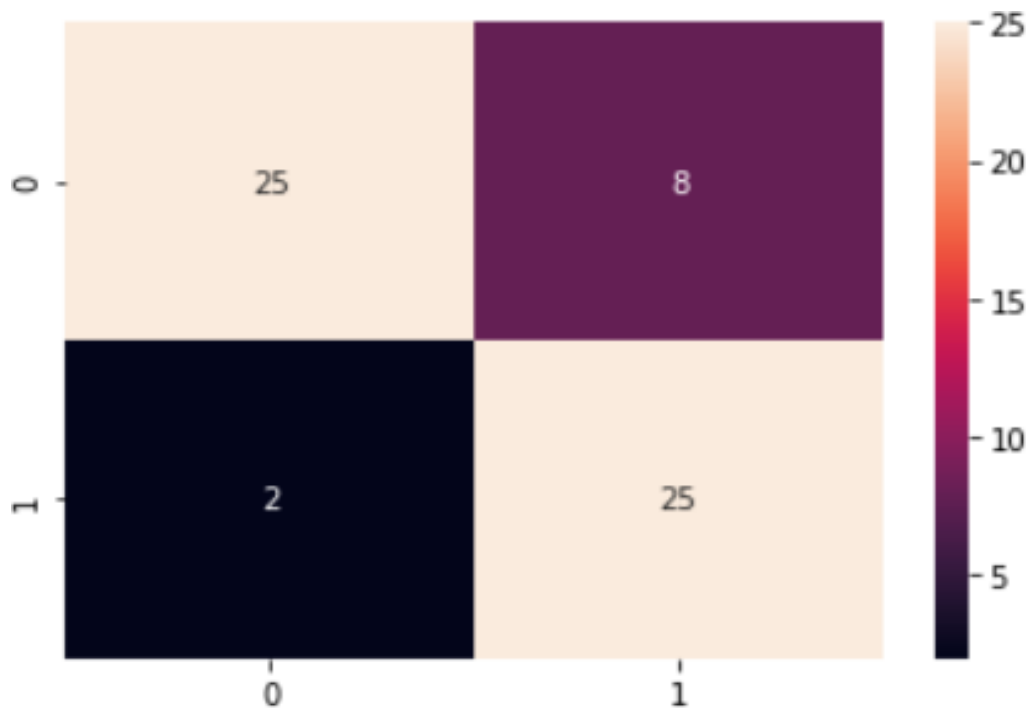


Fig 27

$$T_p = 25$$

$$T_n = 25$$

$$F_p = 2$$

$$F_n = 8$$

$$\text{Accuracy} = \frac{25+25}{25+25+2+8} = 0.833$$

$$\text{Precision} = \frac{25}{25+2} = 0.926$$

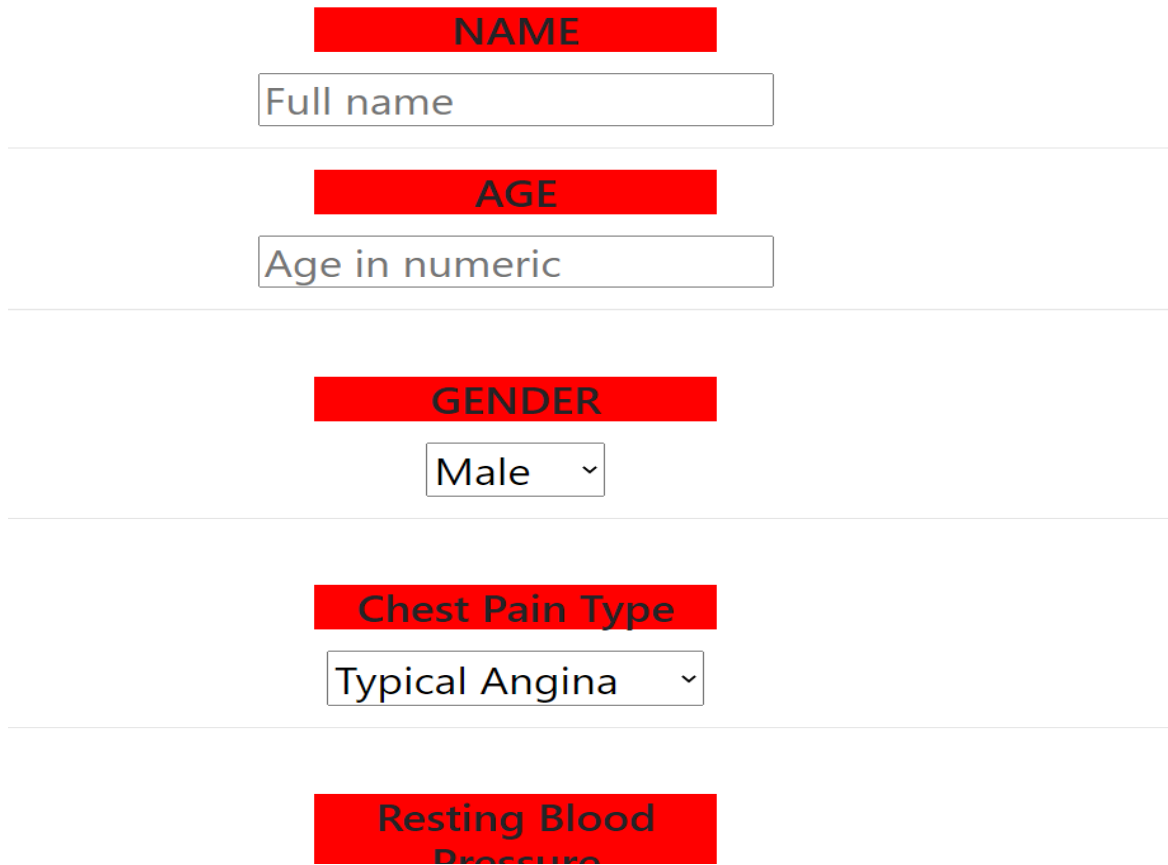
$$\text{Recall} = \frac{25}{25+8} = 0.757$$

Algorithm	True Positive	True Negative	False Positive	False Negative	Accuracy	Precision	Recall
Logistic Regression	22	25	2	11	78.333%	0.917	0.667
KNN	23	26	1	10	81.667%	0.958	0.697
Random Forest	24	24	3	9	80%	0.889	0.727
SVM	23	25	2	10	80%	0.920	0.697
Decision Tree Classifier	21	18	9	12	65%	0.700	0.636
Gaussian Naive Bayes	25	25	2	8	83.333%	0.926	0.757

Table 2

From the above table, it is clear that the highest value of accuracy, precision and recall is achieved through **Gaussian Naive Bayes Algorithm**. Therefore, it is best suited for prediction.

4.2 USER INTERFACE



The figure shows a user interface form with five sections, each separated by a horizontal line. Each section has a red header bar and a corresponding input field below it. The sections are: 1. NAME: A red bar with the text 'NAME' in white, followed by a text input field containing 'Full name'. 2. AGE: A red bar with the text 'AGE' in white, followed by a text input field containing 'Age in numeric'. 3. GENDER: A red bar with the text 'GENDER' in white, followed by a dropdown menu showing 'Male' with a downward arrow. 4. Chest Pain Type: A red bar with the text 'Chest Pain Type' in white, followed by a dropdown menu showing 'Typical Angina' with a downward arrow. 5. Resting Blood Pressure: A red bar with the text 'Resting Blood Pressure' in white, followed by an empty input field.

Fig 28

- User interface makes it easy for the commoner to use and analyze their results by entering the required parameters.
- Users only interact with the front end of the model

CHAPTER-5 CONCLUSIONS

5.1 CONCLUSION

As the number of deaths from heart disease increases, it is important to create a system for predicting heart disease effectively and accurately. The motivation for this study was to find the most efficient ML algorithm for detecting heart diseases. This study compares the accuracy scores of the algorithms such as decision trees, logistic regression, random forest, and naive Bayes algorithms for predicting heart disease using UCI machine learning repository datasets. The results of this study show that the Gaussian Naive Bayes is the most efficient algorithm for predicting heart disease with 83.334% accuracy. In the future, you can improve your work by developing web applications based on the Naive Bayes algorithm and using larger datasets compared to those used in this particular analysis. This allows them to produce better results and assist healthcare professionals. Predicting heart disease helps to analyse the disease effectively and efficiently.

The proposed system is GUI-based, user-friendly, scalable, reliable and extensible. The proposed work model also helps reduce treatment costs by providing a timely initial diagnosis. The model also serves as a training tool for medical students, making it available to doctors and cardiologists as a soft diagnostic tool. General practitioners can use this tool for the initial diagnosis of patients with heart disease.

REFERENCES

- [1] Maneesha, T., Hafeez, S., & Krishna, H. (2018). Prediction of Heart Disease Using Machine Learning Algorithms. In *International Journal of Engineering & Technology* (Vol. 7, Issue 2). www.sciencepubco.com/index.php/IJET
- [2] Yazdani, A., Varathan, K. D., Chiam, Y. K., Malik, A. W., & Wan Ahmad, W. A. (2021). A novel approach for heart disease prediction using strength scores with significant predictors. *BMC Medical Informatics and Decision Making*, 21(1). <https://doi.org/10.1186/s12911-021-01527-5>
- [3] Gao, X. Y., Amin Ali, A., Shaban Hassan, H., & Anwar, E. M. (2021). Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method. *Complexity*, 2021. <https://doi.org/10.1155/2021/6663455>
- [4] Rajdhan, A., Agarwal, A., & Sai, M. (n.d.). *Heart Disease Prediction using Machine Learning*. www.ijert.org

SOFTWARE DOWNLOAD REFERENCES:

- **Python:**
<https://www.python.org/downloads/>
 - **Visual Studio Code:**
<https://code.visualstudio.com/>
- Libraries:**
- **Flask:**
<https://pypi.org/project/Flask/>

- **Seaborn:**

<https://seaborn.pydata.org/installing.html>

- **SkLearn:**

<https://scikit-learn.org/stable/install.html>